BULLETIN
OF "CAROL I" NATIONAL DEFENCE UNIVERSITY
https://buletinul.unap.ro/index.php/en/

Researching disinformation using artificial intelligence techniques: challenges

Ștefan Emil REPEDE, Ph.D. Candidate*

*"Lucian Blaga" University, Sibiu, Romania
e-mail: stefan.repede@ulbsibiu.ro

Abstract

The present article aims to address a series of problems generated by the use of artificial intelligence models for the study of the creation and dissemination of false information beginning from the difficulties in defining and classifying established terms, continuing by exemplifying the way some of the established databases in the research field of fake news are built and ending by noting the differences in their labeling.

Keywords:
fake news; misinformation; disinformation management; natural language processing; NLP; artificial intelligence; machine learning; cyber security.

1. Introduction

Social media was initially seen as an expression of free speech, positive mobilization, and democracy. Currently, studies that include social networks rather categorize these as a threat to democracy (Yerlikaya and Aslan 2020). Social media platforms can be heavily used to manipulate the masses by various types of actors with political, social, monetary, or radical agendas. The issue of media influence on certain agendas emerged mainly after the event known as the Arab Spring and continued in the 2016 US political election, the UK Brexit, the 2017 French Presidential Election, the 2019 Turkish Election, the Pandemic of COVID-19 in 2019, all the way to the major escalation of the Russian-Ukrainian war in 2022. These events have been affected by propaganda campaigns and the spread of disinformation and fake news. The extensive use of disinformation worldwide has emphasized the need to create methods that can flag and combat the use of disinformation.

With the development of emerging technologies such as artificial intelligence (AI), the scientific community proposed various methods involving the use of technologies considered "state-of-the-art" with the aim to combat the phenomenon of creating and disseminating false information. An example of technology studied for the classification of written information from online media, the so-called fake news, was based on the use of Natural Language Processing (NLP) models. The discipline of natural language processing, also known as computational linguistics, is a branch of computer science that uses AI to research written and spoken human languages.

Misinformation and disinformation are classified as the 9th point of interest in the EU Cyber Security Agency's report of October 2022 (ENISA 2022), which states that the use of cloud computing resources, tools, and AI algorithms supports the fabrication of malicious information. The report also states that detecting and minimizing the spread of false information on social media is still among the most important technical approaches to disinformation management. This concept refers to the process of identifying, analyzing, and mitigating the spread of false or misleading information in order to minimize its negative impact on individuals, organizations, and society as a whole. Disinformation management involves steps such as monitoring the target media, detecting potential false information, comparing it with reality (fact-checking), cataloging it and establishing a response or the necessary countermeasures to counter it.

The first step in disinformation management is monitoring and detection (Schia and Gjesvik, 2020). This involves monitoring the spread of information across various platforms and channels such as social media, news websites, and online forums. Automated tools and manual analysis are used to identify potential disinformation campaigns and track their spread. After the detection phase, the next step is fact-checking, which involves checking the accuracy of the information presented.

This step is necessary to determine the appropriate response and countermeasures that should be taken (N. N. Schia 2020). Response and countermeasures involve developing and implementing strategies to counter the spread of the identified campaign or minimize its impact. The measures may include targeted advertising, public messaging campaigns, or requesting the flagging or removal of malicious information from various online platforms (Schia and Gjesvik, 2020).
Managing disinformation is a continuous process that requires collaboration between various stakeholders, including government agencies, media organizations, and civil society groups. Tackling disinformation in general requires a range of skills and expertise from those involved, expertise that includes data analysis, social media monitoring, communications, and public relations strategies. Overall, effective management of misinformation is essential in today's digital age, where the spread of false or misleading information can have serious consequences for individuals, organizations, and society as a whole (US Department of State 2023).

Unfortunately, according to the Global Risks Report 2023 (WEF 2023), disinformation management is a process that is either not initiated or is in its early development stage and its effectiveness is viewed as poor or extremely poor.
Disinformation management is a cybersecurity issue primarily because the spread of false information can be used as a tool to manipulate public opinion or create a false sense of reality (US Department of State 2023).

Fake news – possible definitions and categorizations

An initial problem that arose with the scientific study of the issue related to the creation, dissemination, and consumption of fake news was the definition of the concept. In principle, there are several criteria for the classification of false information, groupings in which the term fake news is only one of the proposed ways of manifestation (Zafarani, Zhou, et al., 2019). The problem imposed by the meaning of the term fake news has been widely discussed in recent years, especially against the background of several campaigns reported in the international media (Baptista and Gradim 2022). The term „fake news" is currently considered inaccurate from a technical point of view, because it describes a wide variety of mass media products, although it is (Gelfert 2018) currently present in the Romanian legislation in article 404 of the Criminal Code (Romanian Parliament 2009), which criminalizes „the communication or dissemination, by any means, of fake news, data or false information or falsified documents, knowing their false nature, if this endangers national security", the article is largely copied from article 168^1 of the version of the Criminal Code of Romania that dates from 1968 (Romanian Parliament 1968).

Speech or news that incites violence is easier to identify than speech that incites hatred, denigrates the rule of law, or slanders certain social groups. The latter is not always clearly identifiable. What is considered unacceptable to one individual may

https://doi.org/10.53477/2284-9378-23-21

be something completely different to another. This kind of difference of opinion creates a certain ambiguity about what constitutes hate speech in a digital context because this medium can include, in addition to outright falsehoods, mistakes in the reporting of facts, opinionated commentary, political satire, or inaccuracies. To disambiguate this often-referred-to concept, several types of classification of false/misleading information have been proposed, which are addressed in the following lines:

2.1. Classification by intent. A proposed classification of fake information, which includes the term fake news, was proposed in the Journal of the NATO Center of Excellence for Strategic Communications (NATO Strategic Communications Centre of Excellence 2020). Understanding the intent behind the campaign/fake news allows for addressing the causes of misinformation and developing prevention, education, or accountability measures. The classification contains several four categories that take into account the intention behind their propagation:

Disinformation – defined as the intentional creation and distribution of false/manipulated information with the intent to deceive/mislead. An example of disinformation can be considered back in 2022 when a false narrative was launched according to which the majority of Romanians want their country to leave NATO and the EU and there is no Romanian political party that can capitalize on this move. It was promoted by a local radio station linked in other situations to promoting misinformation and fake news. The narrative is contradicted by opinion polls (Necşuţu 2022a).

Misinformation – is that false/misleading information that has been distributed without the intention to manipulate or mislead. The main point different from the first type of misinformation representing it is the intention behind its spread. An example of mistaken news coverage occurred on 26 February 2022, when the television channel "Antena 3" mistakenly presented footage from a 2013 video game called Arma 3 as being from Russia's war against Ukraine (Radu 2022).

MALinformation – is a term created by media researcher Hossein Derakhshan, published as a co-author in a Council of Europe report entitled „Information Disorder" (Wardle and Derakhshan 2017) and later adopted by UNESCO. This refers to information that is true and contains correct references, but which is intentionally transmitted negatively to cause actual harm or the imminent threat of actual harm to a person, organization, or country. For example, a post made in the archive titled „Paradise Papers" (Osborne 2017) about the offshore investments of the British Monarchy revealed that many members of the royal family had evasive offshore investments. The campaign was intended to harm the British Monarchy and not to inform the public about their illicit practices.

Propaganda – information, predominantly biased or misleading, that is disseminated with the aim of promoting a cause or political point of view. An example of such a

narrative can be seen in 2022, when the war between Russia and Ukraine is depicted as a war waged by NATO against Moscow (Cezar 2022). The claim about an alleged NATO threat against Russia had circulated long before it was picked up in Romania. It was promoted by Russian propaganda to justify Moscow's appetite for new territories (invasion of Georgia in 2008, invasion of Ukraine in 2014 and 2022) and was based on older Soviet narratives that NATO „encircled" the USSR with its bases. As Russian forces began to claim defeat in Ukraine, the narrative was altered and the new claim is that Russia is actually fighting NATO/the West and that Ukrainians are being used as cannon fodder (Necşuţu 2022b).

Fake news – is information whose falsity is verified and which is intentionally spread. An example of fake news repeatedly circulated in Romanian media claims the Netherlands opposes Romania's accession to the Schengen zone because the maritime port of Constanța threatens the supremacy of the port of Rotterdam. This false narrative was reiterated in 2022 in the context in which Bucharest hoped that Romania would be admitted to the Schengen area by the end of the year (Peiu 2022). This type of news had already appeared 10 years earlier, originally released by the Voice of Russia. The fake news cycle states that the Netherlands will never agree to Romania joining the Schengen area, fearing that the port of Constanța could become the largest port in Europe, thus having an irreversible impact on the Dutch economy which relies heavily on the trade entering and leaving the port of Rotterdam. In fact, the competition between the two ports is out of the question, as the port of Rotterdam is better positioned geographically and has superior infrastructure and operational capabilities (Veridica 2022a).

2.2. Stylistic classification. Another approach to the classification of fake information, which includes the concepts of fake news, disinformation, and propaganda, is focusing on the stylistic way of composing media materials and includes a total number of 5 categories. It was proposed within the American State Library of North Dakota (library-nd.com 2023):

Fake or lying news (false or deceptive). This concept refers to information that is intentionally fabricated or manipulated to mislead readers or viewers (Gelfert 2018). This can include completely fabricated stories as well as news stories that have some element of truth but are distorted or taken out of context to support a particular agenda agendă (Baptista and Gradim 2022). One such example is the online rumor of the death (due to a heart attack) of George Soros on 05/15/2023, originally published on a Twitter account (@PoliticsFAIRL) and picked up by reputable accounts. The claim was not based on any real evidence (LaMagdeleine 2023).

Misleading information. Misleading articles are those that contain partially or completely inaccurate information or that are presented in a way that is designed to mislead readers or viewers (Zafarani, Zhou, et al. 2019). Unlike fake or deceiving news, misleading articles may contain an element of truth, but that truth is being

taken out of context or presented in a way that is designed to promote a particular agenda or point of view (Gelfert 2018). An example of this type of information is the one according to which the reform of the justice system in 2022 will lead to the undermining of the Constitutional Court, Romania will lose its sovereignty, the constitution will no longer be respected and Romanian justice will be conducted in Brussels, to the liking of the West. This misleading narrative was launched in the context of the debates on the justice laws of 2022. In reality, the amendment of the law on the status of judges and prosecutors did nothing but align the Romanian justice system with the European one, respecting the principle of the supremacy of European law (Veridica.ro 2022b).

Polarizing or biased content (slanted/biased). Polarizing or biased content refers to news articles or reports that are presented in a way that favors a particular point of view or agenda (Schia and Gjesvik 2020). The content that falls into this category is not necessarily fake. The news reports true information but does so in a biased manner. This type of partisanship can be political, ideological, or cultural and can manifest itself in various ways, including selective reporting, sensationalism, or the use of loaded language (Baptista and Gradim 2022). Polarizing content may reflect an entity's desire to sway readers' opinions in a certain direction, or it may reflect its attempt to create a memorable news story. Examples related to this type of content occur when only news featuring a certain ideology/political party is presented by a news channel and the others are ignored. Similarly, a political party can only be presented negatively. Another example of polarizing media can be given by an advertisement that supports unproven scientific data (Drew 2023).

Manipulated/modified data. This concept refers to text, images, or video that has been intentionally altered or edited in a way that misrepresents the original content. This may include the use of manipulated images, edited videos, or selective quotes taken out of context (Zafarani, Zhou, et al. 2019). This data is usually used to create false narratives or to support a particular agenda (Zellers, et al. 2019). A telling example of this type of fake information is the video recording created by „deepfake" technology in which the president of Ukraine asked his countrymen to surrender to Russia (The Telegraph 2022).

Humorous pieces of media (including all its forms such as satire, parody, or jokes) (Baptista and Gradim 2022). Such news is intentionally fabricated or exaggerated for comic effect (Figueira and Luciana 2017). Unlike fake or misleading news, humorous news is not intended to mislead or deceive, but rather to entertain. Satirical news may use humor to comment on social or political issues or to expose absurdity or hypocrisy (Gelfert 2018). Even though these types of news are not intended to deceive, they can sometimes be mistaken for genuine news, especially if they are shared out of context or without proper attribution (Schia and Gjesvik 2020). An example of satire is the news story published by the US media group The Onion, known for its humorous news content, „Jimmy Carter wins boxing match against

Jake Paul". This news story, while obviously fake due to the age difference between the former US president and the social media personality, contains an edited photo of the two and the content of the article is presented in the style of a boxing gala recap (the ONION 2023).

2.3. Classification according to impact and motivation. A classification that aims to be more exhaustive and takes into account the motivations behind the creators of the fake as well as indications of the possible impact that each type of fake information can have if it is distributed in an enabling environment has been compiled by the European Association for Viewers' Interests (EAVI) and contains a total of 10 categories classified according to impact (neutral, small, medium, large) and motivation (monetary, political/power, humor/entertainment, passion/ extremism or misinformation) (EAVI 2022):

Propaganda – used by governments, corporations, or nonprofit organizations to control attitudes, values, and disseminated information. This can be beneficial or harmful depending on the motivation behind the campaign which can be created to support a state policy or to create a negative sociopolitical state (neutral impact and motivation related to politics and passion)

Clickbait – offers sensational headlines that attract attention but are misleading because they do not reflect the written content of the material (low impact, motivated by money and entertainment value). Examples of clickbait headlines can be: "You won't believe...", "X things you need to know...", "A weird trick...", "This is what will happen if... ", "The best X...". Examples of this type of news are mostly found in the fashionable sections of the tabloids: "What does Bianca Drăgușanu say about her second child. "I have everything I need, for sure (Lixandru 2023)." From the point of view of disinformation, it is relevant that this type of headline can be used to spread campaigns containing false information.

Sponsored content – has a similar form to that of editorials but disguises ads without making it clear to consumers (low impact and monetary motivation). This type of advertising is considered harmful, especially among young people who cannot differentiate between disguised advertisements (sponsored content) and online normal news articles (McAlpine 2019).

Satire and Hoax – is a humorous social commentary that varies in quality and may have a subtle meaning (low impact and humorous motivation). This class is similar to the humorous pieces in the media (chapter 2.2) with the specification that usually these contents are polarizing/biased in favor of certain agendas.

Error – news or information containing false facts as a result of involuntary errors (the impact is reduced and the motivation is based on misinformation). This type of error can perpetuate false information by not verifying the original source.

Partisan - news that claims to be impartial, includes interpretations of the facts, has an ideological factor, and includes only the facts that confirm a position or policy, ignoring the others (ideological motivation and average impact). This type of news is used in propaganda campaigns to increase the level of credibility of the narratives presented. In such cases, genuine experts or pseudo-experts may be invited to support a point of view but claim impartiality. Partisanship was used during the 2016 US presidential election to so-called impartially create a positive image of one of the candidates (EAVI 2022).

Conspiracy theory – news that simplistically explains complex events as a response to fear or insecurity. These cannot be scientifically verified and the data that denies the respective theories is considered evidence that actually confirms the hypothesis (high impact, ideological motivation, or related to misinformation). Such theories were used during the COVID-19 pandemic to link the vaccine to 5G technology and the Microsoft corporation to create an anti-Western attitude.

Pseudoscience – news that supports theories such as miracle cures, the anti-vaccine movement, and misrepresents real scientific studies with exaggerated or false claims. (high impact, political or monetary motivation). Pseudoscience was used throughout 2020-2021 to promote various anti-EU narratives. Thus, a narrative taken from the eastern zone stated that the anti-Covid measures decided by the authorities are ineffective. The campaign to contest the sanitary measures taken to combat the SARS-CoV-2 pandemic continued in 2021 and pleaded for alternative treatments that did not receive the approval of the Romanian or European health authorities (Arbidol, Ivermectin) that were promoted by obscure doctors or influencers without medical training, at the same time contesting the effectiveness of anti-Covid vaccines employing pseudo-scientific data - either false information or some taken out of context. Of note during this period was the focus on adverse reactions to vaccines, realized through different medical professionals, usually with specializations that have nothing to do with virology, respectively by so-called experts in alternative medicine (Gomboș 2021).

Misinformation – includes a mix of real and fake information combined with false associations, processed content, and misleading headlines. Even if it aims to inform, the author does not know that the information used is false (high impact and motivation are to misinform). Disinformation usually involves concentrated action by different entities and has a clear purpose behind it, usually ideological or military. A famous disinformation campaign took place in World War II when the British government discovered the radar and did not want to tip the enemy about this fact. Therefore, they started a media propaganda campaign in the UK which created the myth that eating carrots helps develop night vision and the UK pilots benefit fully from this discovery (Smith 2013).

Bogus – content that is completely fabricated and spread with the obvious intent to misinform. This category may include guerrilla marketing tactics, software bots,

or fake comments. This content is intended to bring financial gain or political/ ideological influence (high impact, political or monetary motivation). For example, according to Meta's Q1 2023 Quarterly Adversarial Threat Report (Meta 2023), the company removed 40 Facebook accounts, 8 pages and one group for violating Meta's coordinated inauthentic behavior (CIB) policy. The identified network had its origin in Iran and mainly targeted Israel, Bahrain, and France, countries where it operated by posting probably fake ads to gain an increased level of authenticity and insert into established thematic forums on Facebook, Twitter, YouTube, or Telegram. The network could then have been used for various political or pecuniary purposes.

2.4. Lack of consensus. The presented classifications touch upon the problem of identifying and presenting false information from several perspectives. They contain common elements but do not completely overlap. Although the rankings of false information presented started from various differentiations such as intention, impact, and motivation, overlaps of meaning could not be avoided, but no universal classification and implicitly a categorical definition of different types of false information may result from them. Moreover, a unanimous definition is probably not a possibility in the near future due to the multitude of characteristics and diverse forms of manifestation that the creation and distribution of false information has.
The term fake news, although not found in all classifications, which is one of the most used in research and involves the categorization of false information obtained from the public space because it also involves verifying their veracity, is worth noting.

3. Classifying false information and creating datasets for research

The lack of a consensus regarding the definition of fake news and the multitude of sociocultural situations that give perspectives to this phenomenon is stated as one of the initial problems that need to be taken into account in the context of the use of artificial intelligence in the research of the creation and spread of false information. This difference in cataloging creates a problem that, although not obvious, becomes essential in the process of training AI models and testing them. The problem is given by the lack of data ready to be labeled with fake news. To solve this shortcoming, different online projects created data sets for research purposes. Such pre-labeled fake news datasets are collections of news where each piece of information, typically a news article or headline, has been labeled by a human operator as "true" or "false" (or in a variant with more tags) (ISOT Lab 2017). These datasets are used to train and evaluate machine learning models that can automatically classify news articles or headlines as true or fake based on patterns and features learned from labeled data (McIntire 2020).

Such datasets can be used to tune pre-trained language models, such as the Bidirectional Encoder Representations from Transformers (BERT) NLP model

(Ozbay and Alatas 2020), which have already been trained on large amounts of general text data. This can help the models adapt to the specific characteristics of the fake news dataset and improve their performance in identifying fake news (Devlin, et al. 2019).

The fake and real news datasets were constructed by research teams from real-world news articles that were previously vetted by media professionals (Wang 2017; Ahmed, Traore and Saad 2017). Such datasets are used to train and test different automatic fact-checking methods using AI models (Ozbay and Alatas 2020). The article will present some established data sets compiled from news stories that have been previously verified and labeled accordingly.

With the rise of disinformation campaigns, various fact-checking organizations or groups have come together in line with the goal of educating the public to discern true from false information presented by the media. Such examples may include PolitiFact, factcheck.org, or Snopes. Facebook even built its own „International Fact-Checking Network" (IFCN 2016) which has over 90 signatories from all over the world, including Romania.

Such fact-checking entities provide various types of fake news classification methods or annotations to record media verification. For example, PolitiFact takes current news from various outlets or media sources, verifies their claims using official sources or statistical data, and then assigns easy-to-use labels that show how true a media article is, using labels such as: mostly true, true, half true, mostly false, false, pants on fire (PolitiFact 2017); Snopes uses a similar tag code: true, mostly true, mostly false, false, outdated, scam, unproven (Snopes n.d.); the Romanian news verification organization, factual.ro, uses labels such as: true, partially true, truncated, false, impossible to verify (Factual 2016). All these groups base their claims on the percentage of true versus false information in the analyzed media article. It is relevant that fact-checking teams use only open-access sources during these checks.

Developing a fake news dataset typically involves a process of collecting, annotating, and validating news articles from various sources. An overview of the steps involved in creating a dataset of fake and real news from tagged news presented by media outlets includes collecting news articles from a variety of sources, including both traditional news outlets and alternative news sources (IFCN 2016). Next, articles should be checked for fake news content (Kaggle 2018). This is done by expert human annotators who review articles following a pre-set and transparent methodology. After identifying fake news articles, they are annotated with tags indicating whether they are real or fake (Factual 2016) by the same annotators. Finally, the dataset must be validated to ensure that it is reliable and accurate. This can be done by comparing annotation results with other data sources, such as fact-checking websites or other expert sources (Preda, et al. 2022).

By collecting news already tagged by these platforms, researchers involved in fake news studies are able to generate large datasets of certified fake and real news content without having to personally verify and tag them. The basis for labeling a news story is mainly based on the degree of false data that can be found in it. Because there is not yet a standardized way to label false averages and to address certain limitations that arise from using multiple labels for datasets, most established datasets use a binary classification for data, „False" and „True" and ignore labels such as mostly true, half true, or unverifiable.

An example dataset widely used in fake news research called ISOT (Ahmed, Traore and Saad 2017) contains over 1.2 million news articles in various languages, including English, Spanish and Portuguese, and comes from various fact-checking sources, covering a wide range of topics and fields. The dataset was created by the Information Security and Objects Technology (ISOT) Research Group at the University of Victoria in Canada (Ahmed, Traore and Saad 2017). Each news article in the dataset has already been checked by expert human annotators and is labeled as real or fake. Additional metadata such as source, author, and publication date are also provided within the data collection. The dataset is freely available for download and can be accessed via the ISOT Research Group website (ISOT Lab 2017).

A similar dataset used is PolitiFact's LIAR dataset (Wang 2017). PolitiFact is a nonpartisan fact-checking website that evaluates the accuracy of statements made by politicians, public figures, and other prominent individuals. It was founded in 2007 by the Tampa Bay Times, a newspaper in Florida, and has since expanded to include partnerships with other news organizations. PolitiFact rates statements on a „Truth-O-Meter" scale that uses categories ranging from „True" for real news stories to „Pants on Fire" for stories that contain no element of truth or have a nonsensical theme. The site provides detailed explanations and evidence for its ratings and aims to promote transparency and accountability in public discourse by helping people separate factual information from misinformation and propaganda. The dataset called „Liar, Liar, Pants on Fire" is verified by PolitiFact and consists of 12,836 statements made by politicians, which are labeled with six different labels: pants-on-fire[1], fake, barely true, half-true, mostly true, and true. The dataset was created to support research on automated fact-checking and fake news detection by being organized as a database containing the news or statement, information about the source of the statement, the person who made the claim, the context in which it was made, and fact-checking label assigned by PolitiFact (Wang 2017). It has been used to train and test machine learning algorithms for fact-checking and fake news detection and has been cited in numerous research studies (Wang 2017).

[1] "Pants-on-fire" - colloquial expression from the English language that comes from the nursery rhymes "liar, liar, pants, on fire!" meaning someone has been caught lying.

Another established dataset was compiled by KAGGLE and is known as the Kaggle Fake News Dataset (Kaggle 2018). It consists of a collection of news articles labeled either „Fake" or „Real" based on their accuracy and credibility. Kaggle is a popular platform in the field of data science and the organization of competitions involving machine learning technologies. The training dataset contains 20,800 news articles, and the test set contains 5,200 news articles, including a mixture of real news articles from reputable sources and fake news articles from unreliable sources. Fake news articles were collected from various sources on the internet and tagged by people based on their veracity. Real news articles have been collected from reputable news organizations and verified for accuracy by fact-checking organizations (Kaggle 2018). The dataset includes information about the title, text, author, and publication date of each article, as well as metadata such as the source URL and number of social media shares. The Kaggle Fake News dataset has also been used in numerous research studies and machine learning competitions and has helped advance the development of automated systems for detecting fake news.

Datasets such as those presented in this chapter allow researchers to fine-tune their models, focus on achieving superior performance, and more precisely define the classes that an AI model should consider when performing tasks related to identifying fake news, and misinforming content.

4. Using fake news datasets in research. Binary or multiclass approach

The data sets presented in the previous chapter allow the research approach from several perspectives: binary (true vs. false) or multiclass (true, partially true, neutral, etc.) In the process of automatic detection of fake news, a binary approach is a method where the machine learning algorithm is trained to classify news articles into two categories: fake or real (Kaliyar 2021). In contrast, a multiclass approach is a method where the algorithm is trained to classify news articles into more than two categories, such as partially true, completely false, and true, impossible to classify, true, satire, propaganda, etc. (Wang 2017).

A binary approach is commonly used in automated fake news detection because it simplifies the problem and makes it easier to manage. Instead of classifying news into multiple categories, a binary approach only requires distinguishing between two classes: real news and fake news (Chen, et al. 2017). This can help improve the accuracy of the classifier. In addition, as exemplified in previous chapters, the dataset was usually composed of only proven fake and proven real news and labeled by creators only in the two classes to avoid further interpretations (Kaggle 2018; ISOT Lab 2017). By choosing this binary approach the existing tagged data set can be used without the need to create any additional tags. This saves time and resources and allows researchers to focus on improving model performance on the two classes

of interest. In addition, a binary approach provides a clear and easy-to-understand result that can be useful to end users (Kaliyar 2021). For example, a news aggregator or social media platform may only need to know whether a piece of content is fake or not to decide whether to display it to users. A binary classifier can quickly provide this information, making it more useful for such applications (Ahmed, Traore and Saad 2017).

In addition, a binary approach can also simplify the evaluation of model performance. Established metrics to measure a model's effectiveness such as accuracy, precision, regression, and F1 score (another metric that measures model accuracy, combined with regression score) are easier to calculate and interpret when dealing with only two classes (Gnanambal, et al. 2018), this approach thus helping researchers to more easily compare and select the model with the best performance (Yerlikaya and Aslan 2020).

Conclusion

This article highlights the problem of managing fake news from the perspective of its research and some of the problems encountered in this regard. An initial problem concerns the identification of a common typology of false information. The initial use, including in legal texts, of the generic term 'fake news' is currently considered insufficient to encompass the entirety of the existing types of false information and is presently retained within the realm of research because its definition enables researchers to categorize a specific type of news conclusively, having undergone verification with no ambiguity about its classification. This type of approach cannot capture everyday reality in all its manifestations but it represents an initial step in the right direction because the resulting software product can, for example, be used in the rapid initial flagging of certain categories of false information or disinformation campaigns. The paper contributes to the existing literature on automated fake news detection by providing a framework for understanding and addressing this complex phenomenon. We hope that this paper can inspire research and innovation in this field, as well as become a source of information for policymakers and practitioners involved in disinformation management.

References

Ahmed, H., I. Traore, and S. Saad. 2017. "Detection of online fake news using N-gram analysis and machine learning techniques." *International conference on intelligent, secure, and dependable systems in distributed and cloud environments* 127-138. doi:https://doi.org/10.1007/978-3-319-69155-8_9.

Bahad, Pritika, Preeti Saxena, and Raj Kamal. 2019. "Fake News Detection using Bi-directional LSTM-Recurrent Neural nETWORK." *Procedia Computer Science* 165: 74–82.

Baptista, J.P., and A. Gradim. 2022. "A Working Definition of Fake News." *Encyclopedia* 632-645. doi:https://doi.org/10.3390/encyclopedia2010043.

Cezar, Nicholas. 2022. „De ce Războiul din Ucraina nu poate avea învingători." *Național.* https://www.national.ro/politica/de-ce-razboiul-din-ucraina-nu-poate-avea-invingatori-762950.html.

Chen, W., X. Xie, J. Wang, B. Pradhan, H. Hong, D. T. Bui, and J. Ma. 2017. "A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility." *Catena* 151: 147-160. http://dx.doi.org/10.1016/j.catena.2016.11.032.

Devlin, J., M. W. Chang, K. Lee, and K. Toutanova. 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.*

Drew, Chris. 2023. "35 Media Bias Examples for Students." https://helpfulprofessor.com/media-bias-examples-for-students/.

EAVI. 2022. "Beyond Fake News – 10 Types of Misleading News." https://eavi.eu/beyond-fake-news-10-types-misleading-info/.

Emre, Celebi M., and Kemal Aydin. 2018. "Unsupervised Learning Algorithms." doi:https://doi.org/10.1007/978-3-319-24211-8.

ENISA, European Union Agency for Cybersecurity. 2022. "ENISA Threat Landscape 2022." https://www.enisa.europa.eu/publications/enisa-threat-landscape-2022.

Factual. 2016. https://www.factual.ro/.

Figueira, A., and O. Luciana. 2017. "The current state of fake news: challenges and opportunities." *Procedia Computer Science* 121: 817-825. doi:https://doi.org/10.1016/j.procs.2017.11.106.

Gelfert, A. 2018. "Fake news: A definition." *Informal logic* 38 (1): 84-117. doi:https://doi.org/10.22329/il.v38i1.5068.

Gnanambal, S., M. Thangaraj, V. T. Meenatchi, and V. Gayathri. 2018. "Classification algorithms with attribute selection: an evaluation study using WEKA." *International Journal of Advanced Networking and Applications* 9 (6): 3640-3644. https://oaji.net/articles/2017/2698-1528114152.pdf.

Gomboș, Cătălin. 2021. „România 2021: Top FAKE NEWS & DEZINFORMĂRI demontate de Veridica." https://www.veridica.ro/stiri-false/romania-2021-top-fake-news-dezinformari-demontate-de-veridica.

IFCN, International Fact-Checking Network. 2016. "Verified signatories of the IFCN code of principles." https://ifcncodeofprinciples.poynter.org/signatories.

ISOT Lab. 2017. "ISOT Fake News Dataset." https://www.uvic.ca/ecs/ece/isot/datasets/fake-news/index.php.

Kaggle. 2018. "Fake News Detection." https://www.kaggle.com/jruvika/fake-news-detection.

Kaliyar, R.K., Goswami, A., and Narang, P. 2021. "FakeBERT: Fake news detection in social media with a BERT- based deep learning approach." *Multimedia Tools and Applications* (80): 11765–11788. doi:10.1007/s11042-020-10183-2.

LaMagdeleine, Izz Scott. 2023. "No, George Soros Is Not Dead." *Snopes.* https://www.snopes.com/fact-check/george-soros-is-not-dead/.

library-nd.com. 2023. https://library-nd.libguides.com/fakenews/categories.

Lixandru, Livia. 2023. „Ce spune Bianca Drăgușanu despre al doilea copil. «Am tot ce îmi trebuie, cu siguranță»." *Libertatea.* https://www.libertatea.ro/entertainment/ce-spune-bianca-dragusanu-despre-al-doilea-copil-4566539.

Magdin, Radu. 2013. „Constanța: locul unde se ciocnesc interesele. Internaționale." https://cursdeguvernare.ro/constanta-locul-unde-se-ciocnesc-interesele-internationale.html.

McAlpine, Kat J. 2019. "Most people can't tell native advertising apart from actual news articles, according to new research." https://www.futurity.org/sponsored-content-real-news-1961062/.

McIntire, G. 2020. "Fake News Dataset." https://github.com/pmacinec/fake-news-datasets/tree/eb85398bab558791c9f879e9f96ce72a471d2cc9.

Meta. 2023. "Quarterly Adversarial Threat Report Q1 2023." https://about.fb.com/wp-content/uploads/2023/05/Meta-Quarterly-Adversarial-Threat-Report-Q1-2023.pdf.

NATO Strategic Communications Centre of Excellence. 2020. "Defence Strategic Communications ." *Academic Jurnal Volume 8* (8). doi:DOI: 10.30966/2018.RIGA.8.

Necșuțu, Mădălin. 2022a. „Dezinformare: Majoritatea românilor vor ieșirea țării din NATO și UE." https://www.veridica.ro/dezinformare/dezinformare-majoritatea-romanilor-vor-iesirea-tarii-din-nato-si-ue.

—. 2022b. "Disinformation: The West is fighting Russia using Ukraine as proxy." https://www.veridica.ro/en/disinformation/disinformation-the-west-is-fighting-russia-using-ukraine-as-proxy.

Osborne, Hilary. 2017. "Revealed: Queen's private estate invested millions of pounds offshore." *The Guardian.* https://www.theguardian.com/news/2017/nov/05/revealed-queen-private-estate-invested-offshore-paradise-papers.

Ozbay, Feyza Altunbey, and Bilal Alatas. 2020. "Fake news detection within online social media using supervised artificial intelligence algorithms." *Physica A: statistical mechanics and its applications* 540. doi:https://doi.org/10.1016/j.physa.2019.123174.

Peiu, Petrișor. 2022. „Blocadă olandeză la porțile castelului Schengen. Pericolul naționalismului lipsit de inteligență." *Gândul.* https://www.gandul.ro/opinii/blocada-olandeza-la-portile-castelului-schengen-pericolul-nationalismului-lipsit-de-inteligenta-19860233.

PolitiFact. 2017. https://www.politifact.com/.

Preda, A., S. Ruseti, S. M. Terian, and M. Dascalu. 2022. "Romanian Fake News Identification using Language Models." doi:DOI: 10.37789/rochi.2022.1.1.13.

Radu, Cristina. 2022. „Antena 3 a prezentat din eroare imagini dintr-un joc video din 2013 ca fiind din războiul Rusiei împotriva Ucrainei." *Libertatea.* https://www.libertatea.ro/stiri/antena-3-a-prezentat-din-eroare-imagini-dintr-un-joc-video-din-2013-ca-fiind-din-razboiul-rusiei-impotriva-ucrainei-4005144.

Romanian Parliament. 1968. „Codul Penal din 21 iulie 1968 (**republicat**)." https://legislatie.just.ro/Public/DetaliiDocument/38070.

—. 2009. „Codul Penal din 17 iulie 2009, Legea nr. 286/2009." https://legislatie.just.ro/Public/DetaliiDocument/223635.

Rashkin H., Choi E., Jang J.Y., Volkova S., and Choi Y. 2017. "Truth of varying shades: Analyzing language in fake news and political fact-checking." *Proceedings of the 2017 conference on Empirical Methods in Natural Language Processing, EMNLP.* 2931-2937. doi:10.18653/v1/D17-1317.

Schia, N.N., and L. Gjesvik. 2020. "Hacking democracy: managing influence campaigns and disinformation in the digital age." *Journal of Cyber Policy* 5 (3): 413-428. doi:https://doi.org/10.1080/23738871.2020.1820060.

Smith, K. Annabelle. 2013. "A WWII Propaganda Campaign Popularized the Myth That Carrots Help You See in the Dark." *Smithsonian Magazine.* https://www.smithsonianmag.com/arts-culture/a-wwii-propaganda-campaign-popularized-the-myth-that-carrots-help-you-see-in-the-dark-28812484/.

Snopes, Snopes Media Group Inc. fără an. https://www.snopes.com/. Accesat 2 februarie 2023.

the ONION. 2023. "Jimmy Carter Wins Boxing Match Against Jake Paul." https://www.theonion.com/jimmy-carter-wins-boxing-match-against-jake-paul-1850487520.

The Telegraph. 2022. "Deepfake video of Volodymyr Zelensky surrendering surfaces on social media." https://www.youtube.com/watch?v=X17yrEV5sl4.

US Department of State. 2023. "Disarming Disinformation: Our Shared Responsibility." https://www.state.gov/disarming-disinformation/.

Veridica. 2022a. "Fake news: The Netherlands opposes Romania's Schengen accession because the port of Constanța threatens the supremacy of the port of Rotterdam." https://www.veridica.ro/en/fake-news/fake-news-the-netherlands-opposes-romanias-schengen-accession-because-the-port-of-constanta-threatens-the-supremacy-of-the-port-of-rotterdam.

—. 2022b. „Fake news: The reform of the justice system leads to the undermining of the Constitutional Court and Romania losing its sovereignty." https://www.veridica.ro/en/fake-news/fake-news-the-reform-of-the-justice-system-leads-to-the-undermining-of-the-constitutional-court-and-romania-losing-its-sovereignty.

Wang, W. Y. 2017. "«Liar, Liar Pants on Fire»: A new benchmark dataset for fake news detection." *Proceedings of the 55th annual meeting of the association for computational linguistics* 422-426. https://arxiv.org/abs/1705.00648.

Wardle, Claire, and Hossein Derakhshan. 2017. "Information Disorder: Toward an interdisciplinary framework for research and policy making." https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html.

WEF, World Economic Forum. 2023. "The Global Risks Report 2023." https://www3.weforum.org/docs/WEF_Global_Risks_Report_2023.pdf.

XIA, Xin, and LO, David. 2018. "Feature Engineering for Machine Learning and Data Analytics." *Feature* (CRC Press) 335-358. https://ink.library.smu.edu.sg/sis_research/4362 .

Yerlikaya, Turgay, and Seca Toker Aslan. 2020. "Social Media and Fake News in the Post-Truth Era." *Insight Turkey 22.2* 177-196.

Yuandong Luan, and Shaofu Lin. 2019. "Research on Text Classification Based on CNN and LSTM." *International Conference on Artificial Intelligence and Computer Applications (ICAICA).* doi:https://doi.org/10.1109/ICAICA.2019.8873454.

Zafarani, Reza, Xinyi Zhou, Kai Shu, and Huan Liu. 2019. "Fake News Research: Theories, Detection Strategies, and Open Problems." *KDD '19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.* doi:https://doi.org/10.1145/3292500.3332287.

Zellers, R., A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi. 2019. "Defending Against Neural Fake News." https://rowanzellers.com/grover.

Zhou, Z., H. Guan, M. M. Bhat, and J. Hsu. 2019. "Fake news detection via NLP is vulnerable to adversarial attacks." doi:https://doi.org/10.48550/arXiv.1901.09657.